DOCUMENT RESUME

ED 091 669 CS 001 127

AUTHOR Venezky, Richard L.
TITLE Testing in Reading: Assessment and Instructional Decision Making.
INSTITUTION ERIC Clearinghouse on Reading and Communication Skills, Urbana, Ill.; National Council of Teachers of English, Urbana, Ill.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE May 74
CONTRACT NE-C-0-72-4636
NOTE 41p.
AVAILABLE FROM National Council of Teachers of English, 1111 Kenyon Rd., Urbana, Ill. 61801 (Stock No. 05121, $0.95)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Criterion Referenced Tests; Educational Accountability; *Evaluation Techniques; *Guidelines; *Reading; Reading Diagnosis; Reading Improvement; Reading Programs; *Reading Tests; Standardized Tests; Testing; *Test Interpretation

ABSTRACT

This booklet is designed to provide guidelines for testing in reading and suggestions for using the test results in ways which will most benefit the student. Ten canons are presented which are intended to serve as guidelines for program-related assessment within a framework of instructional decision making. They are concerned primarily with the amount and types of assessment which individuals should receive within the bounds of reading instruction, but they also address themselves to some of the problems related to program assessment and to the distribution and protection of assessment results. Some of the canons are: "The function of assessment in reading is to aid in instructional decision making." "The value of an assessment is measured in terms of its unique contribution to a decision." "The content of assessment should be compatible with the content of instruction." "The exactness of assessment should be determined by adaptability of instruction." "The amount of assessment an individual receives within a program should be proportional to his needs within that program." "Program related assessment should provide continual information for making program improvements." "Assessments related to program outcomes should be based upon realistic expectancies." (WR)

Testing in Reading

Assessment and Instructional Decision Making

Richard L. Venezky

ERIC Clearinghouse on Reading and Communication Skills
National Institute of Education

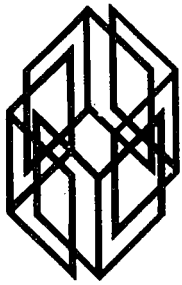
nal Council of Teachers of English

ERIC Clearinghouse on Reading and Communication Skills
National Institute of Education

National Council of Teachers of English
1111 Kenyon Road, Urbana, Illinois 61801

Testing in Reading

Assessment and Instructional Decision Making



Richard L. Venezky
University of Wisconsin

ISBN 0-8141-0512-1
NCTE Stock Number 05121

Published May 1974
ERIC Clearinghouse on Reading and Communication Skills
and
National Council of Teachers of English
1111 Kenyon Road, Urbana, Illinois 61801
Printed in the United States of America

Contents

Foreword

The National Institute of Education (NIE), recognizing the gap between educational research and classroom teaching, has charged ERIC (Educational Resources Information Center) to go beyond its initial function of gathering, evaluating, indexing, and disseminating information to a significant new service: information transformation and synthesis.

The ERIC system has already made available—through the ERIC Document Reproduction Service—much informative data, including all federally funded research reports since 1956. However, if the findings of specific educational research are to be intelligible to teachers and applicable to teaching, considerable bodies of data must be reevaluated, focused, translated, and molded into an essentially different context. Rather than resting at the point of making research reports readily accessible, NIE has now directed the separate ERIC Clearinghouses to commission from recognized authorities information analysis papers in specific areas.

Each of these documents focuses on a concrete educational need. The paper attempts a comprehensive treatment and qualitative assessment of the published and unpublished material trends, teaching materials, the judgments of recognized experts in the field, reports and findings from various national committees and commissions. In his analysis he tries to answer the question, "Where are we?"; sometimes finds order in apparently disparate approaches; often points in new directions. The knowledge contained in an information analysis paper is a necessary foundation for reviewing existing curricula, planning new beginnings, and aiding the teacher in *now* situations.

This booklet, as the title suggests, is designed to provide guidelines for testing in reading and suggestions for using the test results in ways which will most benefit the student.

Bernard O'Donnell
Director, ERIC/RCS

Preface

Nationwide attention is currently focused on accountability and assessment at all levels in our schools, with support in many states from legislation designed to speed implementation. The decade of the seventies has already been deemed the "Accountability Era" and in many ways parallels education's "Efficiency Era" of the early 1900s.

The reading accountability thrust of most school districts includes three basic requisites in order of priority. First, a statement of objective outcomes for reading instruction is required. Second, learning conditions leading to these outcomes must be specified. And third, measurement instruments must be selected to evaluate the degree to which the objectives have been achieved.

Unfortunately, the press of accountability frequently results in a reverse in the priority order of the three requisites identified above. This is due in large part to a simplistic view of reading assessment relying primarily on "objective measurement" by achievement tests. There is a very real risk that the objectives of measurement instruments will come to determine the objectives of instruction. On the other hand, professional educators must recognize the importance of reading assessment and its potential value in relationship to instructional decision making. Therefore, it is essential that public school decision makers and consultants from colleges and universities better understand the assessment role in the instructional process.

Accountability and Reading Instruction: Critical Issues, a 1973 publication of the Commission on Reading, recognized the importance and impact of reading assessment. The Commission membership realized, however, that a vital need existed to present the assessment issue in greater depth. Richard L. Venezky was thus commissioned to develop a monograph designed to provide a set of guidelines "for deciding when, how, and with what to test, and what to do with the results." As Venezky indicates, the monograph's primary goal "is to relate assessment—not just testing in the formal sense—to decisions which are made in the selection,

implementation, and evaluation of reading programs." Commission members unanimously agree that the monograph constitutes a first step toward this goal. Reading specialists, school psychologists, principals, and teachers involved in the development or selection of reading programs will find Venezky's discussion explicating his ten canons of distinct value in furthering their understanding of reading assessment and its relationship to the process of instructional decision making.

Robert B. Ruddell, Director
NCTE Commission on Reading

Introduction

The Origin of Tests

Assessment problems appear to be more severe in reading than in other curricular areas, yet relative to the history of testing procedures, formal assessment instruments for reading were developed late. According to DuBois (1966), psychological testing was an invention of the Chinese, who as early as 2200 B.C. were administering tests periodically to public officials. By the beginning of the Chan dynasty in 1115 B.C., candidates for civic offices were examined formally on a variety of basic abilities, including writing, arithmetic, and archery. The ancient Greeks also used formal tests for civic positions, but this practice is not evidenced in Europe again until 1791 A.D., when the French introduced testing for civil service candidates, a practice which Napoleon abolished not very many years later.

But today's psychological tests owe only an indirect debt to civil service testing. They derive more directly from work done in the second half of the nineteenth century on measuring higher mental abilities and especially on detecting and measuring mental deficits. From an interest in a humane treatment for insane and mentally retarded persons grew a need for uniform classification techniques and, hence, for replicatable assessment procedures. The modern psychological testing movement developed out of this need, principally through the work of Sir Francis Galton, an English biologist who was interested in heredity and especially in the characteristics of related and unrelated persons. By 1905 the Binet-Simon Scale, which used for the first time the concept of mental age, was in use in the Paris schools for selection of subnormal children (Anastasi, 1961, pp. 5-20).

The testing movement which originated in Europe was brought to the United States by Cattell in the 1890s and was translated into educational practice by Thorndike, whose handwriting scale, first presented to the American Association for the Advancement of Science in 1909, was the

first important testing instrument introduced into the public school curriculum in the United States.[1] Six years later the first standardized reading test, Gray's Oral Reading Paragraphs, was published,[2] and within a few years at least ten other oral and silent reading tests were commercially available. In the period from 1914 to 1915, almost two thirds of all published reports on reading were concerned with the standardization and application of reading tests. Today there are well over 150 published reading tests covering every conceivable type and level of reading and every technique of testing which has ever been observed or imagined.

The Use and Misuse of Tests

At the first Invitational Conference on Testing Problems, started through the initiative of directors of state educational testing programs, the main topic of discussion was the misuse of tests in educational contexts. Three recommendations were approved at the end of the meeting, the first of which reads as follows:

> Those conducting state testing programs must recognize that it is their major responsibility to educate teachers and school administrators to a wiser and more efficient use of test materials in dealing with the individual pupil. This is a responsibility which transcends that of insuring high quality in the technical materials and services required in the testing program, *since the techniques of testing have already been developed to a point far beyond the ability of the schools to make the most of the possibilities now presented.* (Anastasi, 1966, p. 4)

How to implement this recommendation is considered today to be a pressing problem, yet the recommendation was not made this year or last year or even ten years ago, but in 1936! And this was long before the development of many of the test construction and data analysis techniques commonly used today.

For selecting test items, ascertaining reliability and validity, standardizing tests, and doing almost any other statistical task related to test development, an extensive literature and a full range of carefully debugged and well-documented computer programs are available. But for using tests intelligently and efficiently within an educational context, and specifically within the context of reading instruction, we are not much better off today than we were in 1936.

In fact, since 1936 the public concern over the potential misuse of tests has grown, as evidenced by such events as the special mention given to psychological tests in the two different Congressional investigations of testing held in 1965.[3] In addition, there is pressure today from minority groups to change or eliminate most standardized testing in education,

based on the conviction that the tests themselves are biased in favor of "Establishment" children.

For reading, the issues related to assessment have become increasingly complex in recent years. Individualization through the use of reading management systems has increased dramatically the amount of assessment which occurs within some reading programs and has thereby created a concern for the proper balance between assessment and instruction. Instruction patterned on skill hierarchies has brought into question the utility of norm-referenced tests and has replaced them in most instances by criterion-referenced measures. Statewide assessments of reading achievement have led to an increasing awareness of the context-sensitivity of test scores, that is, of the relationship between reading ability and a host of other variables, including parent's income level, the amount of money spent by the school on each child, and the competence of the reading staff. These concerns have led in turn to an increasing concern for the proper interpretation of test results and of the problems raised by the unqualified dissemination of test results.

One can now find throughout the United States many instances of proper test use within reading programs, but the overall situation remains far from acceptable. Parents are generally not informed of the significance and utility of their children's test scores, teachers are often inundated with test results for which they have no use, and administrators often invest large sums of money in testing programs without having clearly defined plans for using the test results.

What is lacking is a usable set of guidelines for deciding when, how, and with what to test, and what to do with the results. This monograph is a modest attempt to fill this need for elementary and secondary level reading assessment. Its primary goal is to relate assessment--not just testing in the formal sense--to decisions which are made in the selection, implementation, and evaluation of reading programs. Its audience is school psychologists, reading specialists, principals, and teachers who are involved in the design or selection of reading materials. Much of what is presented here has the ring of plain common sense--those eternal verities that our grandmothers always told to us--but common sense is precisely what is found lacking most often in discussions of testing. One can, for example, marshall an impressive defense for the selection of almost any particular standardized reading test. There are generally available, with the aura of Sinaitic writ, conclusive figures on reliability, validity, and (for newer tests) applicability to different subject populations. But try to find help in answering the ostensibly simple question of whether the test will allow you to make a more intelligent decision than you could make on the basis of already available information. The question is actually a difficult one, and you will find little help, because the answer depends upon very basic considerations of goals, costs, and instructional resources.

Terminology

Distinctions in usage among the terms *testing*, *assessment*, and *evaluation* are neither obvious, well-settled, nor easy to state. *Assessment*, for example, has been evidenced in print in the general sense of "estimation or evaluation" from at least 1626, according to the *Oxford English Dictionary*, yet the latest edition of the *Dictionary of Education* (1973) still does not recognize any use of *assessment* outside of "property valuation" and "counseling." *Evaluation* has been defined in a narrow sense by Tyler as "... essentially the process of determining to what extent the educational objectives are actually being realized by the program of curriculum and instruction" (1950, p. 69). Cronbach, on the other hand, expands this definition to the "... collection and use of information to make decisions about an educational program" (1963, pp. 231ff).

In lieu of a more stable usage to draw upon, I have attempted to adopt definitions for this paper which help to clarify the matters discussed here and yet are not at odds with educational trends.

Testing will be used in this paper in the narrow sense of a formal, critical examination.

Assessment, which is perhaps a little pedantic [Strunk (1959) always cautioned against big words where little ones did just as well], will apply to any procedures—including testing—for gathering performance data. Although assessment in reading always centers on performance of individuals, the resulting information is often used for making decisions about a program—thus the slightly misleading phrase *program assessment.* In a strict sense, we do not assess reading programs; instead, we assess individuals and, from the resulting information, make inferences about the program. This latter process will be called *evaluation. Evaluation*, therefore, will be used here in the narrow sense preferred by Tyler, to determine the agreement between program objectives and program achievements. To reduce potential ambiguity in the use of *assessment* and *evaluation*, the phrase *program-related assessment* will be used when the emphasis is on the individual assessment problems of program evaluation.

Evaluation is program-oriented and backward looking. It asks how much a program achieved relative to a set of achievement expectancies. In contrast, assessment is individual-oriented and usually forward looking. It asks how well an individual can perform relative to the types of instruction which are available to him. In this sense, assessment is diagnostic, and its utility is not in predicting or proving, but in improving.

The ten canons which follow are intended as guidelines for developing program-related assessment within a framework of instructional decision making. They are labeled canons not in the sense of ecclesiastical law, nor in Kant's sense of fundamental a priori principles, but in the simpler and

more general sense of aphorisms or axioms which govern the scientific treatment of a subject. They concern primarily the amount and types of assessment which individuals should receive within the bounds of reading instruction, but they also address themselves to some–but not all–of the problems related to program assessment and to the distribution and protection of assessment results. For the aid of the reader who is not acquainted with the terminology of testing, a brief glossary is included in the appendix.

1

The function of assessment in reading is to aid in instructional decision making.

To understand the full implications of this canon, we must agree on two principles: first, that time and money for instruction are scarce commodities in any school program, and second, that *assessment is not instruction.* Whether an assessment requires fifty seconds or two hours, and whether the cost per child for the assessment involves "merely" the teacher's time or extra costs for tests and analyses, there is some real allocation of time and expense for all assessments. The absurd case–which very frighteningly is becoming more probable every day with the onslaught of the reading management systems–is when the entire school day is devoted to testing. Under such extreme conditions, the real costs of testing are obvious.

On the second assumption–that assessment is not instruction–there is always the hedge, "But assessment is important for instruction." True, but that's not the issue. Many things are important for instruction, proper facilities and a hearty breakfast included, but facilities and food are not instruction either.

Some may believe that testing builds character or that it teaches children to cope with the realities of the twentieth century. I doubt that these beliefs are true, but I don't want to argue them because they too are irrelevant to the current issue, which is simply that no instructional goal is realized by assessment alone.

There is an instructional goal–and only one goal–which *may* be realized or aided through assessment, and this is instructional decision making. Knowing a child's reading level or his competence in pronouncing

initial consonant clusters or his ability to select the main point of a story from four options is of little value unless these data are collected as an aid to a particular decision. Collecting periodic data on reading ability, as many schools and school districts do today, merely for "knowing" what is happening is a monstrous waste of time and money and serves no purpose other than to create suspicion among politicians and parents.

To sanction periodic testing solely for the purpose of curiosity, that is, solely for knowing what is happening, is to encourage inefficient and irresponsible management. Educators have a greater responsibility toward instruction than just curiosity. It is expected that they will engage in continual evaluation and modification and consequently will plan program assessment as an integral component of curricula. There is a need to evaluate the effectiveness of programs, as will be discussed in later canons, but there is no merit to the accumulation of test data for the sole purpose of having the scores. To argue that such data are potentially valuable for some as yet undefined research and therefore should be continually collected and archived is to perpetuate a myth about scientific discovery. Progress in science does not result from the gradual accretion of data into larger and larger repositories. Instead, it comes—as the noted historian Karl Popper has continually stressed—from revolutionary ideas which themselves guide the *selective* collection of data for support.[4] In other words, *there is nothing scientific about a folder full of test scores, unless those scores were gathered with a specific set of decisions in mind.* Reading programs are built around instruction. Often, but not always, there are decisions which can be made during the instructional program—whether or not a child needs further work in a particular skill, which reading materials might be most interesting and challenging to a particular student, and so on. Assessment *aids* in making such decisions, but the need for the decision must be established before the assessment is done.

That assessment *aids* in decision making brings us to another important implication of Canon 1: assessments *aid* in decision making but they are not the sole basis for instructional decisions. Since Canon 2 addresses itself directly to the contribution which assessment makes in such situations, it is sufficient for the present to point out that for almost every instructional decision there is available a variety of information upon which the decision might be based. What constitutes necessary and sufficient information must be ascertained in relation to the decision itself and especially in relation to the consequences of different types of decision errors.[5]

2

The value of an assessment is measured in terms of its unique contribution to a decision.

First, there must be an instructional decision to be made; then, there must be a decision strategy, that is, a decision-making procedure which defines what information to gather and how to translate this information into a decision. It is with the selection of a decision-making procedure that we are now concerned. Take as an example the distribution of students at the beginning of first grade into different reading groups. Why not spend the first month of school testing each child individually on IQ, readiness, attitude toward school, and a variety of other variables which have been identified as predictors of reading success? One reason for not doing this is that most schools have neither the personnel nor the financial resources for such extensive testing, and another is that most of the test scores would be redundant. But a more important reason is that even if the tests which would be used under this plan were not redundant, they would make only a marginal contribution to the decision-making process. The number of different instructional groups into which students are placed is generally small, and the differences in predictive ability of even the most extensive formal tests over informal teacher judgment have never been shown to be large. Therefore, only a limited amount of formally obtained data is required for an adequate decision. If the children have been in a kindergarten, especially one which uses a prereading program, there may already be sufficient data available for placement.

Determining how much additional data to collect for a particular decision, given a cost per datum, cannot be done accurately for instructional programs. In statistical decision theory, as developed by Wald (1950) for quality control of industrial processes and as extended by Cronbach and Gleser (1965) to psychological tests for personnel selection, the contribution of a test to a decision-making process is measured through a mathematical payoff function. To determine the payoff (or benefit) of a given strategy, one must first know the probability of each outcome and be able to assign relative values to each. For a variety of reasons, not the least of which is our inability to assign such values to educational outcomes, formal decision theory is not directly applicable to assessment in reading. Nevertheless, some of the notions contained in decision theory can be incorporated into informal assessment guidelines. One, which originates before the development of statistical decision theory, relates test validity to the practical effectiveness of a test used for selection

(Taylor and Russell, 1939). If, for example, students were divided at the beginning of first grade into two equal-sized groups on the basis of predicted reading ability and nonformal testing procedures resulted in about 60 percent correct placements, a predictive test with a validity coefficient of .50 could be expected to give 79 percent correct placements, or a gain of 19 percent over the nonformal gain-derived score. In the terms of Canon 2, the unique contribution of the test, assuming its validity were .50, would be an additional 19 percent correct placements. (It is important to note that the value of a test, given high reliability, depends very strongly upon the test's validity.)

In some situations, a test adds nothing of significance to what is already known, and, further, two tests are not necessarily better than one and are rarely, if ever, twice as good as one. Over fifty years ago Gates found that the marginal increase in giving a battery of comprehension tests over a single comprehension test was small. His results, stated in terms of correlations, were: "A single comprehension test given in 3.5 to 30 minutes yields a correlation of 0.7 to 0.8 with a composite of comprehension tests representing from 4 to 8 hours of reading" (1921, p. 462). There is little evidence to indicate that any lower correlations would be found with tests published in more recent years.

Even if exact values cannot be assigned to expected outcomes of a decision, it is important to be aware of the nonlinear relationship between cost and probability of correct decision. The cost of testing, as measured by time and expense, increases much more rapidly than does the probability of finding a "true score." A parallel to this phenomenon can be found in attempting to increase test reliability by increasing the number of test items. Doubling the number of test items will raise a .60 reliability to .75, but quadrupling the number of items gains only 11 percentage points over doubling (from .75 to .86). For these reasons, an application of formal decision theory often leads to the selection of an inexpensive procedure with moderate reliability over a more expensive procedure with higher reliability.

Before formal testing should be undertaken, one should ask whether or not the information already available is sufficient for making the required decision. Consider as an example the typical diagnosis-prescription unit. A pretest indicates who needs instruction in a particular skill. Those so identified are sequenced through instructional materials—worksheets, readings, or whatever—and then, typically, a posttest is given to determine if mastery has been reached. But sometimes the information available from daily assignments allows as accurate an assessment of mastery as does the posttest. In such circumstances, the posttest makes no unique contribution to the decision-making process and therefore is without value.

There are actually two issues intermeshed in deciding whether or not to test. One is whether the information gained will add anything to what is

already known or can be learned through everyday instruction. If the answer is no, then testing is not justified, but if the answer is yes, then a second issue is encountered, namely, does the decision require the added precision? This latter question relates to the decision strategy and is discussed in Canon 4.

The content of assessment should be compatible with the content of instruction.

To test which reading skills have been acquired by third graders during a year of reading instruction, we would naturally use a reading test, not a math test or a geography test. But what kind of reading test? An oral reading test? A test for silent reading comprehension? A test of word attack skills? The answer to this question depends (at least in part) upon the content of the instruction. If instruction concentrated primarily upon word attack skills and oral reading, then an assessment of silent reading ability would probably be inappropriate. Such an assessment might measure transfer of learning, but it would not assess learning directly. For those students who failed, no information would be available from the assessment to determine if failure resulted from not acquiring the skills that were taught or not being able to transfer these skills to new situations.

Tests with the same general name often differ widely in the skills they assess. Part of this difference results from a lack of consensus on what skills underlie the reading process, and part results from a lack of overlap of test developers with instructional program developers. Reading readiness tests, for example, agree on very little among themselves. Some concentrate primarily upon letters and sounds, while others concentrate on more general skills, such as those supposedly tapped by picture naming, shape differentiation, and use of oral language. A readiness (or prereading) program which stressed discrimination of beginning and ending sounds would do both its students and its teachers an injustice by using the Metropolitan Readiness Test for end-of-the-year assessment, while a program which emphasized motor control, number knowledge, and picture vocabulary would be equally remiss in employing the Clymer-Barrett Prereading Battery.

The school has a responsibility for intelligent planning of its total reading program, including the designation of prereading skills which are

considered necessary for entry into the reading program. If an assessment instrument is to be used just prior to the beginning of formal reading instruction to select students who are (or are not) ready for such instruction, then the selection of an instrument other than one which is compatible with the school's designation of prereading skills would be self-contradictory. Yet such contradictions are often made and excused by reference to the high predictive ability of such instruments for later reading success. But schools are charged with the task of improving achievement, not predicting it. A predictive score—whatever its validity (and reliability)—is not useful for instructional decision making unless it is based on precisely those skills which the school has selected for instruction, and if this is true, then the predictive index is of less interest than the scores on specific skills.[6]

In other words, educators, not test designers, should decide on the content of instruction. Once this decision is made, assessment procedures can be planned and assessment instruments designed or selected. These same arguments can be applied, *mutatis mutandis,* to program-related assessments and to assessments for diagnosis and prescription. First comes the instructional program and then the assessment instruments—not vice versa.

The most obvious incompatibilities between assessment and instruction are those in skill selection. More subtle forms of incompatibility can result from the specific paradigms used in different tests. Measurement of reading vocabulary, for example, is done by unaided recall of a meaning or a word or an opposite; by matching, classification, or multiple choice; by sentence completion; or by a host of other techniques which differ from each other by more than visual configuration. Children who have practiced deriving word meanings from context may be at a disadvantage when tested on opposites, and children whose main vocabulary instruction has involved matching definitions to meanings may have difficulties with unaided recall. As Farr points out in discussing assessment of reading vocabulary, "What is important is that the test sample the same behaviors as those developed through the instructional program. This is not teaching for a test, rather it is selecting a test which measures growth toward the specified objectives of the reading program" (1969, p. 36).

In opposition to this concern for appropriate test paradigms is the view that if a child knows something well, he will be able to demonstrate it in any reasonable situation. Two things need to be pointed out in considering this view: first, assessments are as valuable for those on the lower end of the performance scale (that is, those who don't know something well) as for those on the top, and second, it is not clear what a "reasonable situation" is in relation to assessment. All test paradigms involve a task variable. At the lower grade levels, where assessment is more frequently used than elsewhere and where children are less familiar with testing

paradigms, the task variable could easily be as important for determining test performance as the test content.

Assessment procedures do not necessarily have to match instructional procedures identically; however, the skills assessed must be identical to those taught. If, for example, children are instructed in basic word attack through procedures which always work from the printed word to sound, then a word recognition test which requires the child to select a printed word to match an orally presented one would be inappropriate because it requires an ability which was not taught, namely spelling. Obviously some spelling ability is acquired in reading activities, but an assessment procedure, unless it is specifically designed to uncover incidental learning, should be selected on the basis of what was overtly taught.

4 *The exactness of assessment should be determined by adaptability of instruction.*

What is meant by Canon 4 is that if a program allows for four different instructional procedures, an assessment which classes students into ten different ability levels is too powerful and potentially wastes resources. Canon 4 is a special case of Canon 1; its concern, however, is with the degree of differentiation which assessment should attempt to achieve. Like Canon 1, it is based on the assumption that the best assessment procedure is the one that achieves the assessment goals with the least expenditure of time and money. Every assessment has a cost, but the more exact we require an assessment to be, the higher the cost.

Almost all standardized survey and diagnostic tests for reading produce finer classifications than can be utilized for instructional decision making. This is not to say that such precision is always wasted; there are research needs which demand such differentiation. But instructional needs would be better served by cruder instruments which required less student time for administration and less administrative time for grading and reporting. If two reading tests are equivalent in all respects except for total number of items, then it can be assumed in general that the longer test is the most reliable of the two, that is, is capable of making finer distinctions among ability levels. Therefore, the decision to choose one test over the other can be based on the degree of differentiation which can be utilized in

instruction. If the instructional program cannot take advantage of the number of distinctions potentially obtainable from the longer test, then there is little justification for selecting it over the shorter test. Reliability can be viewed as an indicator of the amount of uncertainty attached to an obtained score. The true score for a subject lies, for a specified level of uncertainty, within a band which has the obtained score at its midpoint. The width of the band relates inversely to the test reliability; the higher the reliability, the smaller the band of uncertainty, and hence the greater the number of non-overlapping bands which can be obtained. Thus, for classification purposes tests of different reliability may be equally useful, or, in other words, the most reliable test is not always the best test.

One attempt to report obtained scores in relation to test reliability is found in the use of percentile bands, based on the standard error of measurement for the test. This procedure is employed, for example, by both the Sequential Tests of Educational Progress (STEP) and the School and College Ability Tests (SCAT). Scores are expressed in terms of a band which extends approximately one standard error of measurement on each side of the obtained percentile position; there is, therefore, about a 2:1 chance that the student's true score falls within this range. However, this process affects the reporting of test results and not the size of the test itself.

But it is not just the fineness of score differences which needs to be questioned in fitting assessment differentiation to instructional adaptability; the utility of score differences for decisions other than assignment of students to groups on the basis of a single variable must also be questioned. Score differences are integral to norm-referenced tests, which are designed to compare individuals to each other through the use of scores obtained from a previously selected normal population. Such tests tell how one child compares to another and how each compares to the normal population, but they do not reveal what a child can do relative to a predetermined criterion or mastery level. Yet it is this latter concern which is at the heart of most instructional decisions which teachers must make.

The alternative to a norm-referenced test is a criterion-referenced test. According to Glaser and Nitko, "A criterion-referenced test is one that is deliberately constructed to yield measurements that are directly interpreted in terms of specified performance standards" (1971, p. 653). Norm-referenced and criterion-referenced tests differ in both content and utility. Writers of test items for norm-referenced tests generally do not begin with a well-defined set of performances which they intend to measure, and further, because of the need to obtain variability among individuals, they avoid items that are either too easy or too difficult, regardless of their importance for the subject matter of the test. Writers of test items for criterion-referenced tests, on the other hand, begin with a definition of desired behaviors, from which test items are derived. Since

variability across the total range of students is not an issue for criterion-referenced tests, the percentage of students who pass or fail an item does not by itself serve for selection or rejection of the item. Finally, norm-referenced tests have their main application in assessing individuals when selectivity is a factor, for example, in determining which students should be placed in a special program which can accommodate only a small number of them. Criterion-referenced tests are applicable when individuals are to be matched to instructional alternatives.

Although the label "criterion-referenced test" is relatively recent, the type of test it designates is not. Drivers tests and Red Cross Life Saving tests are examples of well-established tests which are concerned only with performance of individuals relative to predetermined performance levels. Reading instruction, as it has moved further toward individualization, has tended to place more importance on criterion-referenced assessment. In addition, city- and state-wide assessment programs are beginning to include criterion-referenced tests. Recently, for example, the New York State Education Commission announced plans to develop criterion-referenced tests as a component in its yearly assessment of educational progress in elementary and secondary schools, and the Center for Statewide Educational Assessment reports that "there is a definite trend toward the use of criterion-referenced testing in those states where the results are used for state-wide decision making" (1973, p. 44).

Because criterion-referenced tests are derived from well-defined instructional goals, they are potentially more compatible with instructional alternatives than norm-referenced tests.[7] However, the basic problem of matching assessment exactness to instructional adaptability remains, regardless of the type of assessment selected.

5

The amount of assessment an individual receives within a program should be proportional to his needs within that program.

For an instructional program with a fixed set of goals, the amount of assessment which any student receives should be proportional to his need for what is instructed. That is, the poorer students, as defined relative to a given instructional program, should receive more extensive assessment than the better students. It is important to note that the amount of assessment

is not fixed by an absolute scale of student abilities, but by a relative scale based on each student's abilities relative to a fixed set of academic goals. Thus a student who reads on a fourth grade level might receive more assessment in a fifth grade program than a student who reads on a third grade level and is placed in a third grade program. The negative implication of this canon is that an assessment program which requires the same assessment for every pupil is a poor assessment program. Many of the diagnostic and prescription procedures for teaching reading subskills commit this offense by prescribing that every student should be given every pretest. If in fact there is any sound reason for the selection of subskills in these programs and for their hierarchical arrangements, then mastery of certain "critical" skills must imply mastery of other so-called lower skills which are prerequistes for these skills. Hence, if a student is tested first on one of these critical skills and he shows mastery, then there is no excuse for the insistence that he be tested on each of the lower skills. If, for example, skills A, B, and C are parallel prerequisities for skill D, then it might be best to assess all students on skill D first, and then to assess skills A, B, and C only for those who did not reach criterion on D.

In more general terms, the student who demonstrates competence in actual reading should not be tested on those subskills which the instructional program claims are necessary for the demonstrated ability. Either the skills truly are necessary, which means that his reading ability is a true demonstration of subskill mastery, or some of the subskills are either unnecessary or are not validly assessed. In this latter situation the student might or might not demonstrate mastery of all the subskills. It would be gross mismanagement, however, to force mastery of the separate subskills if indeed the higher skills which they supposedly supported were already mastered. *Subskills should be viewed as means to an end and not as ends within themselves. The primary goal of reading instruction is to teach reading, and therefore no amount of assessment in subskills is substitutable for assessment of reading itself.*

A strategy which is becoming increasingly popular for matching assessment to individual needs involves *sequential* assessment as opposed to *terminal* assessment. In terminal assessment, a single assessment is employed and the results are used for whatever decision must be made. The assessment procedure, whether it be a single test, a group of tests, or a combination of tests and informal observation, must be sufficiently broad to cover the range of abilities which are critical for the decision. In sequential testing, on the other hand, a number of related assessments are used, and the results of each assessment determine what is to be assessed in the next step. Short assessments without high reliability may be used since the borderline cases can always be reassessed.

In an intermediate level reading program, a sequential assessment procedure might work as follows:

Step 1. At the beginning of the year each student is assessed on oral reading, using a 1-2 minute passage drawn from materials which would be encountered after about the first month of school.

Step 2. Children who had no difficulties in oral reading would be tested no further but would be observed informally, depending upon how much differential instruction could be given to these students. Students who did extremely poorly would be assessed with several finer procedures, beginning with basic word attack skills. The middle group would be assessed on oral reading again, but with easier materials.

Step 3. If any uncertainty existed over the abilities of any of the lower group students, even finer assessment procedures would be used, starting with prereading skills assessment.

By this procedure the best readers would have spent only one or two minutes each in assessment, while the poorest may have received as much as an hour each. More importantly, the teacher would have gathered the most extensive data on those students who would require the most special help. Implicit in this as well as in any other valid assessment strategy is the assumption that the type and difficulty level of the actual assessment which is used is selected on the basis of anticipated results. If, for example, a school typically draws from upper-middle-class homes and the children read, on the average, one grade level above some national norm, then oral reading paragraphs should be selected not according to chronological levels but according to the actual reading abilities of the students as observed in previous years. Sequential assessment is especially effective when the abilities of a group are in doubt. A quick, rough test can be applied to ascertain ability ranges and then finer procedures employed for more exact information.[8]

Program-related assessment should provide continual information for making program improvements.

There is an axiom in industry which says that it is inefficient to develop a process just to produce a particular product. Instead, the process must, at the same time that it produces a product, also provide data which can be used to improve the process itself. Educational programs, like industrial

systems, should also be viewed as dynamic entities that can develop over time through continual monitoring and adjustment. An effective reading program is one that not only achieves certain reading objectives but also has built into it the continual assessment which allows for modification and improvement. Whether through diabolic plotting by publishers, through sloth on the part of educators, or through chance positionings of the celestial orbs, reading programs in the past have generally been viewed as black boxes which were accepted *in toto* when in favor and rejected *in toto* when their fortunes waned. The result has not been salubrious for either teachers or children. For the maintenance of complicated aircraft, this black box view has it merits. Malfunctions in specific systems are corrected by unplugging one black box and plugging in another. The defective unit is either repaired or scrapped. In this manner, an expensive aircraft is not grounded while a specific malfunction is repaired. But there are no comparable black boxes for teaching reading, and furthermore, reading programs are seldom all good or all bad, as many people have been led to believe because of the overemphasis on particular opinions concerning initial reading instruction. Notions like "phonics," "linguistic approach," and "whole word method" are not specifications for building complete reading programs. They are labels—of various degrees of vagueness, but all vague—for one of many instructional components of a complete program.

Most of the decisions which must be made in developing a reading program and most of the everyday decisions which must be made in the classroom implementation of a program are in no way aided by reference to one of the so-called reading methods. Such problems as appropriate readiness, assessment, inductive versus deductive explanation, amount of time to devote to specific skills, appropriateness of certain workbook formats to different ability levels, variation in story types, and degree of individualization must be resolved on the basis of general knowledge about children, learning, and reading, not on the basis of whether one is using or believes in phonics, whole-word, or whatever.

What all of this is leading to is the conclusion that a reading program is a complex matter, the success of which is based not on the ultimate truthfulness of any one hypothesis, but on many different hypotheses, opinions, and assumptions. Under such conditions one should be willing to accept the notion of gradual improvement rather than sporadic total replacement. The implication of this conclusion for assessment is that every responsible teacher and every responsible school system should develop assessment procedures for continual monitoring of their reading programs. Many schools do this through a yearly evaluation session in which formal and informal assessment data are pooled with teacher opinions to decide on changes for the next year's reading programs.

Knowing whether or not a reading program is achieving prescribed

educational objectives at a given time is important, but more important for the continual achievement of educational obligations is having a reading program that is well understood by its instructors and adaptable to changes in children's backgrounds and interests. In different terms, "evaluation, used to improve the course while it is still fluid, contributes more to improvement of education than evaluation used to appraise a product already placed on the market" (Cronbach, 1964, p. 236). Educators should keep in mind that those few inner-city schools which have been identified as having succeeded in teaching reading have developed their programs over periods of time ranging from three to nine years, and that these programs are generally composed of a variety of components selected and adapted by each school (see Weber, 1971).

7

Assessments related to program outcomes should be based upon realistic expectancies.

Program-related assessment is required for a variety of purposes. Schools and school districts want to know how well they are teaching reading; federal programs such as Title I require evaluation of program effectiveness, and many state legislatures require or are intending to require evaluation of programs which receive state aid. In all of these situations there is a potential decision to be made, but the farther one travels from the classroom, the vaguer the decision becomes. For the classroom teacher, evaluation of the reading program can be used to decide on adjustments in procedures and materials. For the principal, evaluation of all of the reading programs in the school can be used to pinpoint overall strengths and weaknesses and, therefore, to decide on allocation of resources. In both of these situations, a set of decision alternatives can be defined and used to guide the assessment process.

One alternative for the teacher might be to adopt totally new reading materials for some part of her reading program. If this is an alternative, then she needs to assess those skills which the available materials give differential emphasis to. This implies that before assessment is done, some consideration is given to the alternative materials which can be used, and especially to how they differ. Then, an assessment procedure is selected to provide the required information.

For program-related assessment which is done to meet evaluation requirements of state and federal agencies, many considerations must be given which are beyond the scope of this monograph. While in theory an evaluation requirement provides an opportunity to obtain data for improving a program, in practice such requirements are most often interpreted as a need to demonstrate success rather than to diagnose openly strengths and weaknesses.[9]

In many situations the decision alternatives are simply to continue or not to continue funding. But to make such decisions intelligently, it is not sufficient to consider only assessment scores, such as class norms or average gain scores. These results must be evaluated in relation to a set of realistic expectancies, and it is in the setting of these expectancies or goals for a program that extreme difficulty is encountered. After many years of intensive efforts to raise the reading levels of poorer readers—especially of those from inner-city and other lower socioeconomic backgrounds—there are relatively few success stories to report. In different words, the lack of a significant number of successful reading programs in inner-city schools (in particular) implies a lack of information on what a successful program should or must contain, how much time and effort is required to implement it, and what results should be expected from it. Therefore, before we banish any program to purgatory for failing to raise the reading scores of any group of children to a predefined level, we should ask whether or not the goals which the program attempted to reach were realistic for the time period allowed and the resources allocated.

Most often, goals for federally funded programs are stated in terms of percentile scores or percentage reduction in failure or in some other quantitative measure. But rarely—if ever—is there evidence to show that *any* program could, under the given conditions, achieve these goals. The solution to this dilemma is not to give up in our attempts to improve reading ability, but to stop confusing long-range hopes with year-by-year expectancies.

It is the ultimate aim of any reading program to produce children of high reading ability, however that ability might be defined. Not even in the panglossian logic that so often pervades the reading market has non-achievement been defined as a program objective. But whether or not the ultimate aim can be realized and in what time is rarely known. Instead, it is usually observed that certain children are not achieving at a desired level, and it is assumed, therefore, that the reading program is at fault. The equally illogical next step is to assume that a deployment of a different package of materials will remedy the problem.

Regardless of the long-range program goal, the question of most importance for program evaluation is what to expect after each year of program use. It would be foolish, for example, to wait five years after implementing a new program before making any evaluation of the total

program, simply because the primary goal of the program is to bring the class average up to a certain point at the end of a five-year period. This procedure is akin to the black-box mentality which was rejected in Canon 6. It is probably true that on the average a school—and especially an inner-city school—should expect a good program to require from three to five years after its installation to reach peak efficiency. This implies, therefore, that major improvements in reading ability should not be expected after only a year or two of program use. (On the other hand, it is assumed that reading achievement during the first few years of a new program should not be worse than under the previous program.)

Program-related assessment during the first few years of a new program should be directed toward these goals: (1) establishment of baseline data for an analysis of reading achievement over a number of years (see Campbell and Stanley, 1963, pp. 34-63), (2) collection of diagnostic data for use in tuning and adjusting the program, and (3) determination of the degree to which the program, as specified, was implemented. Once these goals are achieved, the measurement of program results can become a meaningful concern.

8

Assessments related to program outcomes should be interpreted with respect to program implementation and resource allocation.

In agriculture, where many of the techniques now used in education for experimental design and data analysis originated, evaluation of treatment outcomes seldom requires extensive concern for how faithfully the treatment was applied. If the effects of different periods of ultra-violet exposure on crop yield are to be tested, appropriate instrumentation is designed to ensure that the ultra-violet lights go on and off at designated times. Other variables, such as soil composition, humidity, watering, and temperature, are controlled by similar means. Thus, when we read that cucumber vines which received an extra four hours per day of ultra-violet light of a certain intensity yielded no more than vines which received no extra ultra-violet light, we would seldom question whether the treatment group actually received the specified amount of extra exposure.

But in the evaluation of educational programs the degree to which a program is implemented is a serious concern. Classrooms are not

agricultural plots and educational programs cannot be characterized by such specific measures of light, heat, and nutrition. Until recently, little attention was given in program evaluation to the determination of program implementation, even though major differences in the implementation of the same program were frequently observed.

But for any given method of teaching reading it is not at all clear, unless specified in a particular program, what the full-fledged devotee should be doing in all instructional situations. It is important, therefore, that the designers of programs specify what appropriate implementation is and that the evaluators of such programs use systematic procedures to determine whether the implementation has occurred.

Equally important to assessment of implementation is an assessment of resource allocations. Reading programs are complex matters involving physical facilities, materials, instructors, management, assessment, and students. It is not possible to assay accurately the contribution of any of these components to educational achievements, nor has it even been possible to establish how some of these components vary in relation to each other. Good facilities are important for instruction, for example, but how important? One of Harvard's presidents claimed that Harvard would still be Harvard, even if the faculty had to teach in tents. But would P.S. 11 in New York City still be able to maintain its highly rated reading program if its staff had to teach in tents?

If tents seem a little absurd, consider floor space and equipment. Many American school systems, even in the midst of the financial crises in which they are now entangled, still insist upon allocating from 50 to 100 percent more floor space per child than do schools in most other Western countries.[10] Is the extra space necessary for education or is it an extravagance? Similarly, do the tape recorders, projectors, and other electro-mechanical devices contribute to educational goals in relation to their cost?

The answers to these questions are not known in precise, quantitative terms. And perhaps it is not so important to know precisely what the contributions of either facilities or equipment are as it is to realize that these and many other factors are components of any real instructional program and may contribute to its success or failure. Some reading programs work very well under experimental conditions but fail soon after wide-scale implementation due to their excessive demands on school resources. No matter what the initial outcomes of a program are, those who are responsible for instructional decisions must ask whether or not the program can be sustained with the resources which the school is willing to allocate. A program may require extra aides, extra materials, extra teacher time, or teacher abilities which are not immediately available.

Furthermore a program that is successful in one school may not be successful in another, no matter how similar their students may be,

because instructional capacities differ widely. Some contracted programs in reading reached their contractual goals by utilizing more school resources than could be allocated on a long term basis. Reading instruction was taught for more hours per week than was desired and utilized more aides than could be managed on a permanent basis. If only the program outcomes were assessed, then one would be forced to conclude that these programs were successful. But no matter how important anyone feels reading is, a school or school system is forced to allocate its limited resources to achieve a variety of goals and therefore must limit the time, money, and personnel allocated for reading; therefore, it is important to know what resources are required for sustaining a particular program.

But evaluating the instructional resources which a program requires is often quite difficult. Costs can be determined for materials, personnel, and special activities like test scoring, and the amount of time allocated to the program can usually be estimated, but determining instructional capabilities which the program requires is a different matter. Individualized programs in particular tend to place heavy demands on personnel. Teachers often must work in teams, engage in elaborate record keeping, and be prepared to use a variety of different materials and methods. To establish that a particular program is effective is not sufficient information to allow a school to decide whether it should adopt that program. It is necessary to know at what cost it is effective. This is especially true in comparing different reading systems. The system which produces the largest gains or the highest number of masters of particular skills may not be the best program if it requires an excessive allocation of resources. What is *excessive* is a matter for each school or school district to decide.

9

The distribution of assessment results should be limited to those who are prepared to use them; any other access to assessment results should be limited to those who have a right to know about them.

The people who give tests and other forms of assessment are responsible for the distribution and the protection of the results they obtain. In most circumstances an assessment is directly tied to a program option and is of little interest or value beyond the classroom. But with certain types of assessment, including standardized testing, the distribution of results often raises conflicts between individual privacy on the one hand and the need

for feedback to the teacher and for accountability to parents and to the public on the other. The failure of some school systems to resolve these conflicts properly has led to such drastic measures as a bill now pending before a state legislature which would require, among other things, that all test scores be destroyed within a year after they are obtained.

Computerized data banks, and especially those containing personnel records, have accentuated a longstanding concern for the protection of pupil records. Whether declared by specific legislation or not, individual test scores are not a matter of public domain. If an assessment is properly planned, then there exists both a decision strategy and a well-delimited group of persons who will participate in the decision process. Under such circumstances the persons who must receive results are clearly identified. What remains to be resolved is who else should either be given results directly or allowed access to them if he so requests. Should, for example, newspaper reporters be allowed access? Teachers who are not directly involved in the assessment process? Parents?

The answers to these questions, when not mandated by federal or state legislation, or by school board policy, should be derived from a consideration of the secondary needs which assessment results might serve. These, as suggested above, are (a) feedback to the teacher and (b) accountability to parents and to the public.

Teachers, as professional educators, should have access to all information which might aid in instructing their pupils. But no useful purpose is served by sending to a teacher test scores which are not normally applicable to classroom decisions. If, for example, a school board attempts to evaluate several different reading programs through an assessment which involves both subject and item sampling, it is highly doubtful that the results obtained from any one pupil would be helpful to his teacher. Some might argue that no harm is done by giving the scores to the teachers, since they can ignore them if they so choose. A more tenable position, however, based upon observation of how teachers normally react to test scores for which they have no immediate need, is to inform teachers that certain scores are available for their use, clarify what the scores represent and how they might be used, and then allow each teacher to decide whether or not he wants access to them.

The policy advocated here is based on the assumption that there is no virtue to too much data. Whatever assessments are needed for proper implementation of an instructional program should be incorporated into the program itself. The data from other assessments which result from needs external to the instructional program would then stand little chance of providing new information which could affect instructional decisions. In any case, the teacher should be the one to decide whether particular scores are useful to him or not.

In deciding whether or not to release test scores to parents and to the

public, a totally different issue arises, in that each of these groups has a right to know, which is only vaguely defined but is derived at least in part from the notion that schools are accountable for their actions. It is doubtful, however, that accountability is served by the release to parents and to news media of test scores alone.[11] Perhaps the most flagrant abuse of educational responsibility in this regard is the release to news media of standardized test scores which rank schools according to the reading abilities of their students. The immediate—and seldom contested—implication of such scores is that the schools on the bottom of the list do a poor job of teaching reading while those on the top do a good job. In fact, just the opposite could be true—but this, of course, could never be demonstrated from the scores alone.

To answer the question of instructional efficiency, measures of educational gain and of resources expended are needed, but these are not simple quantities to assay and consequently are seldom offered to the public. It is, nevertheless, a gross exaggeration of the efficacy of schooling and a gross simplification of the variables which affect academic achievement to evaluate schools on the basis of standardized test scores alone; consequently, the release of such scores, if done at all, should be accompanied by a carefully worded explanation of their limitations.[12]

For parents, it is more difficult to argue for not releasing test scores; nevertheless, the same precaution stands—test scores are often misinterpreted and, therefore, schools and teachers have a responsibility for transmitting correct and understandable information. This involves not only a decision on how to explain assessment results, but also a decision on which results to explain. For reading programs which require frequent pretesting and posttesting of skills, frequency of reporting scores to parents also needs careful consideration. The availability of test scores should not be considered by itself as a necessary and sufficient condition for sending the scores home.

10

The form in which assessment results are distributed should be fixed by the decisions which they are to aid.

This sounds so simple, so obvious. If children are assessed by criterion-referenced tests at the beginning of second grade to determine

who needs further help in certain word attack skills and who does not, then the most obvious information which the teacher responsible for instruction wants to have is simply who still needs help in what. To report class averages, percentage scores, or anything else in this situation is superfluous and misleading. Such statistics have their place, but in the example cited the teacher's task is quite well-defined: decide who needs help in which skills. But test scores are rarely formatted for decision making, and the practices which can now be observed in reporting test results are excellent demonstrations of the chasm between testing theory on the one hand and test utilization on the other. Considerable effort has been expended on computer programs for test scoring, but rarely do the designers of these programs concern themselves with making the output directly usable by a non-statistician.

Part of the problem is that certain types of test scores are not useful at all for teachers; the other part is that too little thought has gone into test score reporting. The reporting of percentile bands by the STEP and SCAT tests which was mentioned earlier is a small step forward. Some schools have experimented with computer-generated verbal reports of the form "John X is doing *well* in *oral reading, about average for the class* in *comprehension,* and *slightly below average* in *vocabulary*," where the italicized words are slots in a standard form into which evaluations (good, average, etc.) and subjects are inserted according to a teacher-constructed table which converts numeric scores to categories. However, unless there are decisions which the teacher can make based on these results, little is accomplished by this practice other than a marginal esthetic advancement.

Reporting assessment scores to parents brings in slightly different considerations, in that awareness, rather than decision making, is usually the prime concern. In general, assessment scores—if they are reported at all to parents—should be embedded in more general reports which emphasize the teacher's overall evaluation of the student as opposed to his performance on specific assessments.

A Concluding Note

The gist of the ten canons presented here is that assessment, whether done by formal testing or informal observation, is an integral component of any instructional program and is legitimatized by the need to make decisions at various points during the program's use. Assessment, however, is secondary to instruction. It comes into being only when a decision must

be made, and then only when it can provide information not obtainable by less expensive means. Consequently, the people who plan reading programs should also be the ones who plan assessments. Reviews of reading tests, such as those by Buros (1972) and Farr (1969), should be consulted for information on published tests, and testing specialists may be needed for aid in analyzing results, but the more basic questions related to the specification of the forms and contents of assessments must be answered by the program planners. The canons presented here are intended for precisely that purpose—to aid those who develop reading programs to also develop their own assessment specifications. To the degree that this is achieved, the goal of this monograph will be achieved.

Footnotes

1. Thorndike's handwriting scale and many of the early reading tests are discussed in Nila Banton Smith, *American Reading Instruction* (Newark, Del.: International Reading Association, 1965), Chapter 6.

2. The derivation, standardization, and use of the Standardized Oral Reading Paragraphs are discussed in William S. Gray, "Studies of Elementary-School Reading through Standardized Tests," *Supplementary Educational Monographs*, Volume I (Chicago: Department of Education, University of Chicago, 1917).

3. These investigations are discussed in a special issue of the *American Psychologist* entitled "Testing and Public Policy" 20 (1965): 857-992. Of more recent vintage in the outcry against potential abuse in testing is a National Education Association Task Force Resolution (72-44) adopted in July 1973, which states that "The National Education Association strongly encourages the elimination of group standardized intelligence, aptitude, and achievement tests to assess student potential or achievement until completion of a critical appraisal, review, and revision of current testing programs." *Task Force and Other Reports* (Washington, D.C.: National Education Association, 1973), p. 28.

4. Karl Popper, *The Logic of Scientific Discovery* (New York: Basic Books, 1959). This same point has been made by T.H. Huxley: "Those who refuse to go beyond fact rarely get as far as fact, and anyone who has studied the history of science knows that almost every step therein has been made by . . . the invention of a hypothesis which, though verifiable, often had little foundation to start with. . . ." Cited by A. Koestler, *The Act of Creation* (New York: Dell Publishing Company, 1964), p. 233.

5. Canon 1 has been advocated in slightly different terms by Lee J. Cronbach, "New Light on Test Strategy from Decision Theory," in *Testing Problems in Perspective*, edited by A. Anastasi (Washington, D.C.: American Council on Education, 1966), p. 53: "In every practical use of tests, our aim is to make decisions"; and by R. Glaser and J. Nitko, "Measurement in Learning and Instruction," in *Educational Measurement*, edited by R.L. Thorndike (Washington, D.C.: American Council on Education, 1971), p. 625: "The fundamental task of testing and measurement (in education) is to provide information for making basic, essential decisions with respect to education's instructional design and operation."

6. I am assuming here that the prereading skills are selected in relation to the skills which are emphasized in the first years of reading instruction. If schools insist

upon selecting readiness tests on the basis of predictive value, they should at least be aware of the dependence of this predictiveness upon the skills utilized in teaching reading. "Other things being equal, that test will have greater predictive value which measures the aspects of reading abilities that will be given greatest emphasis in reading." A.I. Gates, G.L. Bond, and D.H. Russell, *Methods of Determining Reading Readiness* (New York: Teachers College, Columbia University, 1939), p. 41.

It should also be pointed out, however, that the predictive ability of a test reflects upon only one of three classes of validity – that of criterion validity. It in no way judges either content or construct validity, which for most instructional purposes are more important gauges of a test's merits than criterion validity. (For a discussion of different types of validity, see the American Psychological Association's *Standards for Educational and Psychological Tests and Manuals*, Washington, D.C.: APA, 1966, pp. 12ff.)

7. A series of articles covering the design and use of criterion-referenced tests, including the problem of defining reliability, can be found in W. James Popham, ed., *Criterion-Referenced Measurement* (Englewood Cliffs, N.J.: Educational Technology Publications, 1971).

8. Sequential testing is discussed in Cronbach, 1966, pp. 56ff.

9. The wording of the evaluation provisions in Title I of the Elementary and Secondary Education Act of 1965, because of its undue emphasis on annual measures of educational achievement, was often interpreted as a mandate for setting unrealistic goals. Sec. 205 (a-5) of the act states ". . . that effective procedures, including provision for appropriate objective measurements of educational achievement, will be adopted for evaluating at least annually the effectiveness of the programs in meeting the special educational needs of educationally deprived children. . . ." We could, perhaps, excuse the naive hope for instant success in "meeting the special educational needs of educationally deprived children. . ." as another case of the reckless optimism that accompanied Federal programs in the early and middle sixties, but the mandatory assessment of educational achievement was ill-founded at best.

10. Figures on floor-space allocations per child for public schools in England are given in Guy Oddie, *School Building Resources and Their Effective Use* (Paris: Organization for Economic Co-operation and Development, 1966). Comparable figures for the USA can be gleaned from various reports in the journal *Nations Schools*. On the consideration of costs in curriculum evaluation, see also Michael Scriven, "The Methodology of Evaluation," in *Perspectives of Curriculum Evaluation*, edited by Ralph W. Tyler, Robert M. Gagne, and Michael Scriven (Chicago: Rand McNally and Company, 1967), pp. 81ff.

11. The guidelines adopted by the American Psychological Association for safeguarding and interpreting test scores provide a well-established base for resolving the problems raised here. Principle 6 of the *Ethical Standards for Psychologists* (1963) establishes the confidentiality of test results: "Safeguarding information about an individual that has been obtained by the psychologist in the course of this teaching, practice, or investigation, is a primary obligation of the psychologist." Principle 14 provides guidelines for releasing test results: "Test scores, like test materials, are released only to persons who are qualified to interpret and use them properly. When test results are communicated directly to parents and students, they are accompanied by adequate interpretative aids or advice." A slightly different set of principles was advanced for telling parents about test scores by the *Test Service Bulletin* (No. 54, December 1959, pp. 1ff.):

"(1) Parents have the right to know whatever the school knows about the abilities, the performance, and the problems of their children. (2) The school has the obligation to see that it communicates understandable and usable knowledge." In this regard, the reader is also encouraged to consult D.A. Goslin, "Ethical and Legal Aspects of School Record Keeping," *National Association of Secondary Schools Bulletin* 55 (1971): 119-126.

12. Gene R. Hawes, "Releasing Test Scores: Urgent or Unthinkable?" *Nations Schools* 89 (1972): 41-52, describes several models for what appear to be carefully planned releases of test data to parents and to news media. The California State Education Department, for example, reports each of the following along with its test scores for each district: minimum, maximum, and median teacher salary, average class size, pupil-teacher ratio, non-teaching personnel, general tax refund rate, general purpose tax rate, assessed valuation per unit of average daily attendance, minority enrollment, scholastic ability (i.e., IQ), pupil mobility, rate of staff turnover, instructional expenditures per unit of average daily attendance, and regular average daily attendance. In Virginia the superintendent of schools held a news conference with the first release of test scores in 1971 to explain the testing procedures and the limitations of the test scores. In Tulsa, Oklahoma, and Columbus, Ohio, extensive explanations accompany all test scores which are released, with a strong emphasis on explaining the multitude of factors which might affect achievement. From the information available on these model cases, it appears safe to conclude that a carefully planned information campaign can overcome much of the misunderstanding which normally accompanies the release of comparative test scores. The efforts expended in cities like Tulsa and Columbus to educate the public on the factors which influence academic progress are—even without the test scores—exemplary instances of educational responsibility. For a survey of dissemination policies in statewide assessment programs, see the Center for Statewide Educational Assessment, *State Educational Assessment Programs, 1973 Revision* (Princeton, N.J.: Educational Testing Service), 1973.

References

American Psychological Association. *Ethical Standards for Psychologists.* Washington, D.C.: APA, 1963.

American Psychological Association. *Standards for Educational and Psychological Tests and Manuals.* Washington, D.C.: APA, 1966.

Anastasi, Anne. *Psychological Testing.* Second Edition. New York: The Macmillan Company, 1961.

Anastasi, Anne, ed. *Testing Problems in Perspective.* Washington, D.C.: American Council on Education, 1966.

Campbell, Donald T., and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research.* Chicago: Rand McNally and Company, 1963.

Center for Statewide Educational Assessment. *State Educational Assessment Programs: 1973 Revision.* Princeton, N.J.: Educational Testing Service, 1973.

Cronbach, Lee J. "New Light on Test Strategy from Decision Theory." In *Testing Problems in Perspective,* edited by A. Anastasi. Washington, D.C.: American Council on Education, 1966, pp. 53-59.

Cronbach, Lee J. "Evaluation for Course Improvement." In *New Curricula,* edited by R.W. Heath. New York: Harper and Row, Publishers, 1964, pp. 231-248.

Cronbach, Lee J., and G.C. Gleser. *Psychological Tests and Personnel Decisions.* Urbana, Ill.: University of Illinois Press, 1965.

DuBois, Philip H. "A Test Dominated Society: China, 1115 B.C.-1905 A.D." In *Testing Problems in Perspective,* edited by A. Anastasi. Washington, D.C.: American Council on Education, 1966, pp. 29-36.

Farr, Roger. "Reading: What Can Be Measured?" Newark, Del.: International Reading Association, 1969.

Gates, Arthur I. "An Experimental and Statistical Study of Reading and Reading Tests." *Journal of Educational Psychology* 12 (1921): 445-464.

Gates, Arthur I.; G.L. Bond; and D.H. Russell. *Methods of Determining Reading Readiness.* New York: Bureau of Publications, Teachers College, Columbia University, 1939.

Glaser, R., and J. Nitko. "Measurement in Learning and Instruction." In *Educational Measurement,* edited by R.L. Thorndike. Second Edition. Washington, D.C.: American Council on Education, 1971.

Good, C.V., ed. *Dictionary of Education.* Third Edition. New York: McGraw-Hill Book Company, 1973.

Goslin, D.A. "Ethical and Legal Aspects of School Record Keeping." *National Association of Secondary Schools Bulletin* 55 (1971): 119-126.

Gray, William S. "Studies of Elementary-School Reading through Standardized Tests." *Supplementary Educational Monographs,* Vol. I, No. 1. Chicago: Department of Education, University of Chicago, 1917.

Hawes, Gene R. "Releasing Test Scores: Urgent or Unthinkable?" *Nations Schools* 89 (1972): 41-52.

Koestler, A. *The Act of Creation.* New York: Dell Publishing Company, 1964.

National Education Association. *Task Force and Other Reports.* Washington, D.C.: NEA, 1973.

Oddie, Guy. *School Building Resources and Their Effective Use.* Paris: Organization for Economic Co-operation and Development, 1966.

"On Telling Parents about Test Results." *Test Service Bulletin,* No. 54, December 1959.

Popham, W. James, ed. *Criterion-Referenced Measurement.* Englewood Cliffs, N.J.: Educational Technology Publication, 1971.

Popper, Karl. *The Logic of Scientific Discovery.* New York: Basic Books, 1959.

Smith, Nila Banton. *American Reading Instruction.* Newark, Del.: International Reading Association, 1965.

Strunk, William, Jr. *The Elements of Style.* Revised and enlarged by E.B. White. New York: The Macmillan Company, 1959. Chapter 5.

Taylor, H.C., and J.T. Russell. "The Relationship of Validity Coefficients to the Practical Effectiveness of Tests in Selection." *Journal of Applied Psychology* 23 (1939): 565-578.

"Testing and Public Policy." Special Issue of *American Psychologist* 20 (1965): 857-992.

Tyler, Ralph W. *Basic Principles of Curriculum and Instruction, Syllabus for Education 360.* Chicago: University of Chicago Press, 1950.

Tyler, Ralph W.; Robert M. Gragne; and Michael Scriven, eds. *Perspectives of Curriculum Evaluation.* Chicago: Rand McNally and Company, 1967.

Wald, A. *Statistical Decision Functions.* New York: John Wiley and Sons, 1950.

Weber, George. "Inner-city Children Can Be Taught to Read: Four Successful Schools." Occasional Paper No. 18. Washington, D.C.: Council for Basic Education, 1971.

Glossary

Construct validity–See *validity*.

Content validity–See *validity*.

Criterion-referenced test–A test which attempts to measure whether or not an individual has mastered a particular objective or set of objectives. Such tests are concerned with how well an individual performs relative to particular abilities and not how individuals compare to each other.

Criterion validity–See *validity*.

Item sampling–A testing technique in which item sets for a test are drawn from an item pool in such a way that all items in the pool receive equal exposure, yet the number of subjects receiving identical item sets is a subset of the sample population. (Generally, each item or item set is assigned to a randomly selected subset of subjects.) This technique is especially effective for describing groups, in that the content domain of the test can be broadened without making the testing load excessive for each subject.

Norm-referenced test–A test constructed especially to differentiate among individuals according to their abilities within a subject domain. Raw scores on such tests are interpreted relative to the distribution of scores obtained from a standardization sample, that is, a representative sample of the intended subject population on which norms were obtained. Items for a norm-referenced test are selected for their ability to differentiate among individuals; hence, neither extremely easy nor extremely difficult items are desired.

Obtained score–See *true score*.

Reliability–The consistency or stability of a test, as measured by repeated administration of the test or its equivalent to the same individuals. Reliability is a predictor of the range of variation in an individual's score on a single test due to random factors. A highly reliable test would yield a relatively small range of scores when administered repeatedly to the same individual (assuming no learning effects), while an unreliable test would yield a wider range of scores under these same conditions. On the different sources of unreliability and the different accepted techniques for measuring reliability, see Anastasi, 1961, chapter 5.

Sequential testing–A testing procedure wherein the item or item set selected for administration at any point during testing is determined by the subject's record of successes and failures on previous items or item sets. The advantage to this form

of testing is that greater effectiveness can be gained from each test item compared to *terminal testing*, wherein the complete set of test items (or tests) for a subject are preselected. However, sequential testing is more complex than terminal testing and requires adaptation of testing to each subject.

Standard error of measurement–A measure of the reliability of a test expressed in standard deviation units.

Terminal testing–See *sequential testing.*

True score–The value which would be obtained if a test score were entirely free of error. In other words, it is the score which would result if all non-systematic influences were removed from an observation. The *obtained score*, on the other hand, is the raw or observed score which generally contains a random error component in addition to the component due to systematic factors.

Validity–The degree to which a test measures what it claims to measure. Validity is usually determined in terms of *content validity*, *construct validity*, and *criterion validity*. *Content validity* applies to how well the test content covers the domain which is measured and is generally determined by inspection of the test items, by administration of parallel test forms before and after instruction, and by inspection of the errors commonly made on the test. *Construct validity* applies to how well the test measures a theoretical construct and is generally determined by correlations between the test and similar, established tests. *Criterion validity* applies to how well a test predicts future behavior and is generally determined by correlating test scores with measures of the individual's subsequent performances in the domain which the test purports to predict.

OTHER ERIC/RCS PUBLICATIONS

Black Dialects and Reading– Bernice Cullinan, editor. Examines the complex interrelationships among black dialect, oral language, and reading, and offers teachers practical suggestions based on the most recent research. Introductory essay identifies the issues involved in teaching black children to read standard English. Diagnostic tools for identifying the child's language base include a comparison of beginning reading texts with first graders' actual speech patterns. Other sections provide classroom strategies for teaching oral standard English at the primary, middle, and junior high school levels. 1974 (NCTE and ERIC/RCS). Stock No. 00572. $3.95 nonmembers, $3.75 members.

Miscue Analysis: Applications to Reading Instruction–Kenneth S. Goodman, editor. Goodman explains miscue analysis, which is premised on the fact that errors children make in reading provide specific and general insights about the learners' strengths and weaknesses. Other authors discuss uses of miscue analysis in the classroom and with children from different language backgrounds, as well as its application to writing instructional materials and teacher training. 1973 (NCTE and ERIC/RCS). Stock No. 03677. $2.50 nonmembers, $2.25 members.

The Politics of Reading: Point-Counterpoint– Sister Rosemary Winkeljohann, editor. In his article "The Politics of Reading" (reprinted here), Neil Postman argues that print is no longer the dominant medium of communication in our culture and the fact that the schools are acting as if it were has broad political implications. Eight reading experts, including leaders in professional organizations, reading specialists, a teacher, and a publisher, then respond to Postman's plea for reevaluation of the role of reading instruction in education. The last paper is a response by Postman to his colleagues' arguments. 1973 (International Reading Association and ERIC/RCS). NCTE Stock No. 04131. $2.00 nonmember, $1.80 member.